Profiling the Female Emigrant: A Method of Linguistic Inquiry for Examining Correspondence Collections

Moreton, E

Original citation:
Moreton, E. (2012) Profiling the Female Emigrant: A Method of Linguistic Inquiry for Examining Correspondence Collections. *Gender and History - Special Issue: Gender History Across Epistemologies*, volume 24 (3): 617-646

Publisher: Wiley-Blackwell

DOI: 10.1111/j.1468-0424.2012.01699.x

aptara	GEND	gend_1699	**Dispatch:** 8-18-2012	**CE:** N/A
	Journal	**MSP No.**	**No. of pages:** 30	**PE:** Carmen

Gender & History ISSN 0953-5233
Emma Moreton, 'Profiling the Female Emigrant: A Method of Linguistic Inquiry for Examining Correspondence Collections' Gender & History, Vol.24 No.3 November 2012, pp. 617–646.

Profiling the Female Emigrant: A Method of Linguistic Inquiry for Examining Correspondence Collections

Emma Moreton

Introduction

Mass emigration from Ireland to the United States in the nineteenth century has been examined in terms of its economic, political and social impact on both home and the New World. Drawing on a range of sources such as census information, shipping records and other public documentation, research suggests that during this period there was an increase in migration amongst females, mostly single women in their late teens and early twenties.[1] Knowing that it was unlikely they would ever return to Ireland, the letter was the main method through which these young women kept in touch with loved ones back home.

Over the past few decades there has been a growing interest in the emigrant letter and how this type of source might inform our understanding of social history during the postal era of globalisation. The sourcing, preservation and documentation of emigrant letter collections are growing, and whilst their value as sociohistorical artefacts is generally accepted, finding the best means to exploit such resources is yet to be agreed upon. For David Gerber, emigrant letters have generally been used in one of two ways: to 'provide color and drama in historical narratives, or to document societal-level and group-level generalizations', or as edited collections which 'let the letter-writers speak for themselves, while providing some background information that enables readers to place the [author] in the general societal framework of a certain place and time'.[2] Influential studies such as William I. Thomas and Florian Znaniecki's *The Polish Peasant in Europe and America*, Charlotte Erickson's *Invisible Immigrants: The Adaptation of English and Scottish Immigrants in Nineteenth-Century America*, Kerby Miller's *Emigrants and Exiles: Ireland and the Irish Exodus to North America*, and Walter Kamphoefner, Wolfgang Helbich and Ulrike Sommer's *News from the Land of Freedom: German Immigrants Write Home*, have demonstrated the value in using personal letters to gain a fuller, multi-perspectival understanding of both the complex social processes of emigration (such as push/pull factors and the role of institutions and communities) and the conditions and daily lives of the emigrants themselves.[3] However, Gerber suggests that while ' . . . social historians have been especially skilled

in understanding large categorical social groups – social classes, ethnic groups, religious denominations, and men and women' they have sometimes failed to understand 'the individual self, in relation to other individual selves, [in relation] to the world'.[4] Gerber's own study of British emigrants in America takes an approach that begins with the individual. Viewing the personal letter as an object of study in its own right, Gerber examines how such documents embody relationships, experiences and mental worlds; and more recent studies taking a similar (qualitative) approach, have examined how, through the emigrant letter (or in the case of Anna De Fina and Kendall A. King, the personal narrative), transatlantic relationships are changed and maintained, identities assimilated and narratives constructed and performed.[5]

Such studies, which place emphasis on the individual, seek to get as close as possible to the letter writers' lived experience. The reality of the authors' lives – or the way in which the authors construe their experiences – is revealed through the language contained within the letter. The analyst examines the content of the letter on a semantic level, making sense of its content by placing it within the context of situation (the circumstances in which the letter was produced) or broader still within the context of culture (the societal pressure for the author to perform in a particular way – by writing the letter in the first place and by respecting a particular culture of letter writing when doing so). In most cases the researcher is inferring outwards from the letter, taking the content of the letter to then make claims about what that content means or what it reveals about the context of situation and culture. Intuitively, the conclusions drawn make great sense; however, methodologically speaking, the conclusions are potentially open to criticism as there is a leap of faith and assumption leading from the analyst's intuition to his/her conclusion/point.

The work of Stephan Elspaβ, Nicola McLelland, Marina Dossena and Arja Nurmi and Minna Palander-Collin goes some way to addressing this methodological issue by adopting a bottom-up, empirical approach to studying emigrant letters.[6] Taking as their starting point words and phrases, they then look at how these words and phrases typically behave in sentences, paragraphs and texts, before considering what these linguistic patterns or phraseology might reveal about the situational and cultural contexts in which the letters were produced.[7]

Marina Dossena, for example, examines the use of formulaic as well as dialectal features of language in a corpus of nineteenth-century Scottish emigrant letters to see how such linguistic strategies contribute to, and reinforce, social bonds between author and recipient.[8] (In its most general sense, a corpus is a collection of texts, designed to be representative of the way that language is used in a particular context.) To do this, Dossena looks at a subsection of letters from the Corpus of Nineteenth-Century Scottish Correspondence: forty-two letters (approximately 27,000 words), dating from 1815 to 1892, by thirteen male informants and two female informants.[9] A close qualitative study of the linguistic features characteristic of the letters teased out some interesting findings. Dossena observed that 'involvement strategies' in the openings and closings of the letters were 'mainly dependent on the conventions of formulaic usage' as set out in letter writing manuals of the time. However, she goes on to say that 'within the body of the letter . . . encoders express their psychological proximity to their recipients by means of other linguistic devices', such as the use of Scotticisms (dialect, which, as observed by Dossena, is often employed humorously to stress a 'common cultural background'); visualisations of context (descriptions of people, places and likenesses)

and epistemic modality (words which express certainty/probability, such as *might* and *suppose*, which are used to 'predict the recipient's reactions or the encoder's suppositions about what is going on at home').[10] Although empirical in nature and taking a more bottom-up approach to identifying salient linguistic features across a range of texts, this study is still primarily qualitative and therefore open to the same criticisms previously mentioned. The conclusions resulting from (what are very interesting) observations would have greater strength if it were possible to test their significance reliably. Are these observations typical, unusual, or evenly distributed across different authors, for example? Quantitative investigation and/or statistical tests would help to make claims about the relevance of these observations. Arguably, without quantitative support, it is difficult to appreciate fully the significance of the linguistic features being noticed.

To demonstrate the value in applying statistical measures to test, challenge or support qualitative observations, Nicola McLelland used quantitative methods of analysis to examine the language of nineteenth-century German emigrant men's and women's private correspondence.[11] Referring to twentieth-century studies in language and gender,[12] McLelland argues that research into gender differences has lacked clarity, in part, because it has been qualitative rather than quantitative: '[p]roblems arise when data that are essentially anecdotal in nature are treated as if indicative of general trends without appropriate statistical analysis'.[13] In her study, McLelland focuses on some of the linguistic strategies identified in recent scholarship as being more typical of women in conversation – such as the use of epistemic modality (as previously mentioned, words like *might* and *suppose*), hedging devices (words like *seem*, *believe* and *sometimes*), and question tags (such as *isn't it? shouldn't I?* and *don't they?*) – and then uses statistical methods to test whether such gender differences are evident in a corpus of nineteenth-century letters.

The analysis involved using two corpora: Corpus One (a pilot corpus) containing twenty-two letters by women and twenty-two by men (approximately 30,000 words), dating from 1850 to 1900 and representing seven female and eight male authors; Corpus Two (a much larger, more representative corpus) containing ninety-one letters by men and ninety-one by women (approximately 84,000 words), from the same time period and representing thirty-eight female and thirty-eight male authors.

Of the linguistic features investigated in Corpus One, only the discourse particle '*doch*' (often used as an intensifier or emphatic device) showed any significant difference between genders, being more frequently used by female authors in phrases such as '*ich denke*' (broadly translated as 'I think') to soften assertions. The data also gestured towards female authors being more likely to 'soften imperatives, express more wishes, and [be] more emphatic in their formulations than . . . men', although these findings were somewhat tentative.[14] However, when the same investigation was carried out using the larger corpus (Corpus Two), these findings received little statistical support. What the findings did show, however, was that the female authors used more intensifying adverbs (in English these would be words like 'very', 'really' and 'so'), they were more likely to address the recipient in the body of the letter and they referred to themselves using the first person (I) more frequently than their male counterparts. The data also showed that the female authors tended to adopt more politeness strategies ('*bitte(n)*' (a verb meaning to ask/request) and '*Bitte*' (similar to 'please' or 'you're welcome') when making requests; however, as McLelland points out, this

finding could simply be a result of more requests being made by women than by men in the first place. McLelland hypothesises that the high frequency of '*doch*' in Corpus One may be explained in terms of educational background. It is one particular author who contributes over a third of all occurrences of '*doch*' in Corpus One and this author also adopts a more colloquial, speech-like style in her letters, indicative of a lower level of education.

The study by McLelland (as with similar quantitative studies by Elspaß, and Nurmi and Palander-Collin) demonstrates the possibilities and opportunities of using more quantitative methods of analysis both to examine letters and to understand gender; however, it is a very difficult balance to achieve between offering a rigorous, replicable and systematic quantitative approach whilst at the same time not losing sight of the very personal, idiosyncratic and subjective data at hand, that is, the author behind the letter.[15]

The current article builds on this body of quantitative research, proposing a complementary methodology which is based on the theories and techniques of corpus linguistics for examining emigrant letters – a methodology which attempts to bridge the gap between the content observed and the conclusions that are later drawn from that content; and one which moves constantly between the quantitative and the qualitative and back again. Whilst recognising that linguistic choices will reveal something about the context of situation, context of culture and how the author construes events and perceives the world, corpus linguistics decontextualises the components of language. Corpus linguistics does not look at language at the semantic level, but instead looks at language at the textual level. It is a mode of study that takes language out of its flow and reality, freezing it and rearranging it to give 'new perspectives on the familiar'.[16] It draws on what is known about language and how language is used to make conclusions about how the author is using language. The conclusions drawn are based on empirical data collection: frequencies, distributional patterns and proportions, and because of the design of the corpus, it is possible to move constantly between the individual and the group and back again, noticing what is typical or unusual about one text when compared with many texts. Corpus linguistics, as applied to the study of correspondence, takes language out of context, reorganises it to notice new things based on quantitative investigation, and then puts the findings back into context to try to build a picture of the life and experiences of the author. This approach makes it possible to investigate systematically the language used by different authors and then to notice what those authors each have in common. As such, it provides a multi-layered approach to examining language and gender, allowing the analyst to test whether linguistic observations are about gender alone or gender in combination with other social, cultural or economic factors (such as age, class, location or level of education, for instance).

What is a corpus and what can it do?

A corpus can be defined as a 'bod[y] of naturally occurring language data stored on computers' and corpus techniques of analysis as the 'computational procedures which manipulate this data in various ways . . . to uncover linguistic patterns which can enable us to make sense of the ways that language is used'.[17] The 'body of naturally occurring language' can be anything from a few sentences to a large set of texts (the

term 'texts' here refers to both written language and spoken transcriptions), but the main point to emphasise is that the data has been collected for a specific purpose, with the aim being something 'other than to preserve the texts themselves because they have intrinsic value', which is, as Susan Hunston explains, what distinguishes a corpus from a digital archive.[18] A corpus does not simply preserve and store texts so that they can be accessed more easily and by a wider number of people; rather, a corpus is designed with the intention of being representative of a particular type of text – newspapers, academic essays, letters, political speeches and so on, from a particular era, on a particular subject or by a particular socio-economic group and so on. Representativeness is usually achieved by 'breaking the whole down into component parts and aiming to include equal amounts of data from each of the parts'.[19] So, for example, a corpus of political speeches during the UK general election campaign of 2010 might include an equal weighting of speeches for the news media, TV debates and public addresses, by a range of politicians from the three major parties. What goes into the corpus, then, depends on what the corpus will be used for and what research questions it will seek to address. However, it will also depend on 'what [data] is available', and quite often the analyst is negotiating a fine balance between selecting texts that are representative and working with whatever texts are available.[20] This issue of representativeness is always problematic and arguably no more so than when working with letters. In the case of emigrant letters, the analyst is always working with what is available – designing a representative corpus of personal letters is simply not achievable as there is no way of accounting for the experiences of those emigrants who chose not to write, who could not write or whose letters were lost, destroyed, or, years later, for various reasons, not donated.

The second thing that distinguishes a corpus from a digital archive is the way in which data content is explored and analysed. Although the data in a digital archive may be accessed online, without the need physically to visit a library or an archive, the content is generally studied linearly (as one would do with an original manuscript). Digitisation alone (and by that I mean optical character recognition (OCR) scanning or transcribing the letters and saving them in an electronic format) makes a document more accessible and to a certain extent more searchable (in a very limited sense of the term); however, it does not allow the collection to be explored in depth, or in creative ways. With a corpus, the data is stored in such a way that 'it can be studied non-linearly, and both quantitatively and qualitatively', using computer software.[21] The data (in this case the emigrant letter) can be marked-up in various ways – for contextual information such as gender, age, date of correspondence, socio-economic status, religious denomination, location (home and New World); for key themes such as homesickness, money, work, family, health, American life or for pragmatic features such as apologising, making requests, humour and so on. The data can also be annotated for parts of speech (word classification) and semantic categorisation. This markup and annotation allows individual letters and subgroups to be easily searched and compared in relation to one another and in relation to the whole. Additionally, computer software allows the content of the corpus, or subsections of the corpus (known as subcorpora) to be explored in ways that would be difficult, and in many cases impossible (depending on the amount of data being examined and the type of search being carried out), using more traditional methods of content analysis. Computer software allows the analyst to observe recurrent patterns, distributional trends and other statistical features, which

would be hard to notice through reading alone. For this reason, it is often the data that will lead the investigation, pointing the analyst to features of the texts which they may not have noticed otherwise.

Background remarks

The letters used in this study are borrowed from Kerby Miller (Curator's Professor, Department of History at the University of Missouri), who has created an archive of Irish emigrant correspondence. The collection contains well over 5,000 Irish emigrant letters of which the Lough family correspondence are a small but significant part, having been referred to in several publications.[22] In the early 1950s, a few of the Lough letters were donated initially to Arnold Schrier, an American graduate student, by Canice and Eilish O'Mahony of Dundalk, Co. Louth.Schrier, now Professor Emeritus at the University of Cincinnati, employed these and other letters in his book, *Ireland and the Irish Emigration, 1850–1900*.[23]In the 1970s and 1980s, the rest of the Lough letters were donated to Miller by the O'Mahonys and by Edward Dunne and Mrs Kate Tynan of Portlaoise, County Laois.Both Miller and Schrier, who subsequently collaborated on Irish migration research, made transcriptions of the letters and Miller returned the original manuscripts to the donors.In most cases, Miller's collection contains photocopies of the original manuscripts together with the typed transcripts.

The complete letter collection is described as the LOUGH Corpus. The individual letter series (subcorpora) of each sister are described as: LIZZIE Corpus (which contains letters written by Elizabeth Lough), ANNIE Corpus, ALICE Corpus and JULIA Corpus. All italicised words and phrases are examples taken from the letters. Words in capitals represent the lemma (that is all variations of a particular word form, so BE would represent all forms of the verb to be: is, am, are, was, were etc.). Raw frequencies are presented in angle brackets.

The following is a summary of the background research carried out by Professor Miller, relating to the Lough family. The Lough (pronounced Locke) sisters came from a Roman Catholic family in Meelick, Queen's County, Ireland. There were five sisters – Elizabeth, Alice, Julia, Annie and Mary. Four of the sisters – Elizabeth, Alice, Julia and Annie – emigrated to America between 1870 and 1878, while Mary, the youngest sister, remained in Ireland with her mother and father. Elizabeth (or Lizzie) Lough (later Elizabeth Walsh) emigrated circa 1870–1871 to Winsted, Litchfield County, Connecticut. She worked as a seamstress and housewife, having five children – three boys (Tom, John William and James) and two girls (Alice and Catherine Elizabeth). Her husband, Dan Walsh appears to have died in the early twentieth century, before Lizzie; there is no mention of Lizzie in letters written by her sisters from 1912 onwards, so it is suspected that she died not long after her husband. Alice (or Alisha) Lough (later Alice Elliott) also emigrated circa 1870. She too lived in Winsted between 1876 and 1880, but then moved to Westfield, Hampden County, Massachusetts in 1881 where she remained, as a housewife, until she died sometime between 1918 and 1928. Alice married Edward Elliott and had seven children – five boys (Edward, James, William, John and Phillip) and two girls (Mary Elizabeth and Alice). Annie (or Nan) Lough (later Annie McMahon) emigrated circa 1878. She lived in Winsted from 1884 to 1928 (when her letters stop), working mainly as a servant and housewife. She married John McMahon; she did not have any children. Julia Lough (later Julia McCarthy) was

the last sister to emigrate – in late September 1884. She lived with her sister Lizzie (and husband Dan) in Winsted between 1884 and 1895. She then moved in 1895 to Torrington, Litchfield County, Connecticut, where she remained until 1928 (when her letters stop). Julia was something of a success story, working as a seamstress to begin with, then an apprentice dressmaker, before becoming a professional dressmaker and opening up her own shop that employed several members of staff. She married Thomas McCarthy and had several children although only Elise is named in the letters.

The LOUGH corpus

The starting point for a corpus investigation is quantitative. What is unusual, interesting or typical about a text can only be explained by comparing it against other texts. It is this 'comparative information that quantitative corpus data can provide'.[24] By constantly moving between the cohort and the individual it is possible to notice both what is typical and distinguishing about a text or texts. In this study I will investigate the letters of the four Lough sisters who emigrated from Ireland to the US in the mid to late nineteenth century. I will examine the collection (or corpus) as a whole to see if there are any recurring patterns or phraseology, and what these might reveal; I will also compare the letter series (or subcorpora) of each individual sister to see how their language differs, and what this might reveal. I will, where relevant, use two reference corpora of emigrant correspondence from around the same period: twenty-one randomly selected letters by male Irish authors from a range of socio-economic backgrounds and twenty-one letters by female Irish authors. Letters for the reference corpora were also borrowed (and transcribed) from Professor Miller's archive. These two corpora (although very small for the purpose of this study) will allow me to test whether the findings from the LOUGH Corpus are representative of female emigrant correspondence more generally as well as the extent to which the language of male and female authors differs.

There are a total of ninety-nine letters in the Lough collection held at the University of Missouri; however ten of these were excluded from this study as they did not contain sender information, making it difficult to assign these letters confidently to one of the four subcorpora. As I will be comparing the letter series of each individual sister it is important that each correspondence is correctly assigned to a subcorpus – a wrongly assigned letter could affect the results. Another three letters were discounted, as these – although part of the Lough collection – were not written by any of the four sisters. Although the corpus is relatively small (compared with many corpora, for example, the British National Corpus or the Bank of English, which reach into millions of words), it will nonetheless provide a good foundation on which later studies, looking at larger bodies of data, can build. Corpus linguistics is about making comparisons by looking at what happens in one text and then seeing if this is typical of many texts, and vice versa. The same statistical measures are used when looking at a small amount of data as when looking at a large amount, thereby making it possible to compare corpora or subcorpora of different sizes.

To prepare the letters for corpus analysis they first needed to be digitised and then saved in plain text format (a format which is compatible with most corpus software programmes). The digitised letters are an exact copy of Professor Miller's original transcriptions with all spelling variations, punctuation, omissions, deletions and insertions in keeping with the original manuscript. The process of digitisation and markup,

Table 1: The LOUGH Corpus.

LOUGH Sisters	Number of letters	Number of tokens
Annie	38	18933
Julia	33	12269
Alice	10	3587
Elizabeth (Lizzie)	5	3488
TOTAL	86	38277

Table 2: Breakdown of senders/recipients.

	No. of letters sent			
Addressee	Annie	Julia	Lizzie	Alice
Mother	9	23	1	3
Sister (Mary Lough, later Fitzpatrick)	26	10	0	7
Mother and Sister	1	0	0	0
Nephew – James	1	0	0	0
Niece – Alice	1	0	0	0
Father, Mother and Sisters	0	0	2	0
Father and Mother	0	0	2	0

and the issues and challenges of working with original manuscripts and various versions of transcription, is a topic for discussion in its own right, but these issues, although very relevant, are not within the scope of this article. It was not necessary to mark-up the letters for contextual information (date, location, gender, etc.) for the purpose of this study; nor was it necessary to annotate the letters for parts of speech as this study is intended to be data-led (i.e. basic frequency information will lead the investigation; I will not be approaching the corpus with specific, predetermined grammatical/structural searches in mind). All quantitative findings will need to be examined qualitatively (using concordance lines, which display the words in context) to establish how a word or phrase is functioning – whether as a noun, verb, adjective, etc.

As shown in Table 1, (after removing those letters which cannot be assigned to one of the Lough sisters) the LOUGH Corpus contains eighty-six letters – a total of 38,277 words. Annie Lough, the third sister to emigrate in 1878, appears to have written the most letters of the four sisters – a total of thirty-eight letters (18,933 words) between 1890 and 1928, nine of which were to her mother and twenty-six to her sister Mary (see Table 2), both of whom remained in the Lough's home town – Meelick, Queen's County, Ireland – until their deaths. Julia Lough, the last sister to emigrate in 1884, also wrote regularly – mainly to her mother (twenty-three letters) and also her sister (ten letters) – a total of thirty-three letters (12,269 words) between 1884 and 1927. Elizabeth and Alice were the first sisters to emigrate between 1870 and 1871, yet they wrote the smallest number of letters. Elizabeth wrote five letters (3,488 words) to her mother, father and sisters between 1876 and 1877, when she first emigrated to the US and Alice wrote ten letters (3,587 words) to her sister and mother between 1888 and 1914 (two when she first emigrated and then another three at roughly five-year

Table 3: Type/token ratios for the LOUGH Corpus.

	Annie	Julia	Lizzie	Alice	Total
Type	1718	1463	654	597	2681
Token	18933	12269	3488	3587	38277
Type/Token Ratio	9.07%	11.92%	18.75%	16.64%	7.00%

intervals – five of the letters are not dated, but the content would suggest they were written several years after emigrating). It should be pointed out, however, that this information is based on the number of letters held in Professor Miller's archive (in other words, the number of letters which were donated). As mentioned previously, when discussing the issue of representativeness, there is no way to know how many letters were actually sent or how many were lost or destroyed.

Having grouped the Lough data, it is now possible to explore the content of the letters using computer software. There are a number of useful corpus analysis programmes available, some of which are web-based – Antconc,[25] Wmatrix[26] and Sketch Engine[27] – while others are computer-based – WordSmith[28] and ConcGram.[29] I have chosen to use Antconc for two reasons: first, it is freely available online and second, it has certain functionalities which I am interested in using for this investigation – specifically, the n-gram procedure which will be discussed later in the article.

The starting point – simple frequency data

The first calculation that Antonc can provide is something called a type/token ratio, which can be obtained for the corpus as a whole and for each subcorpus. The term 'token' refers to the total number of words in a corpus. The term 'type' refers to the number of original (or different) words in the corpus. So, for example, the word HOME occurs 193 times in the LOUGH Corpus, which would equal 193 tokens, but would only count as one type.[30] The type/token ratio is calculated by dividing the number of types by the number of tokens – this figure is then expressed as a percentage. A low type/token ratio (tending towards 0%) suggests that certain words are being used over and over again. A high type/token ratio (tending towards 100%) suggests a more diverse range of language is being utilised, with fewer words being repeated. Looking at Table 3, the data shows that the LOUGH Corpus has an overall type/token ratio of 7%. Breaking this down by subcorpora, the data shows that Annie has the lowest type/token ratio (9.07%), followed by Julia (11.92%). The type/token ratios for Alice and Lizzie are slightly higher, 16.64% and 18.75% respectively.

On the surface, this might suggest that Annie's and Julia's letters are more formulaic and repetitive, whereas Alice's and Lizze's letters contain greater lexical variety and complexity. However, it is more likely that this difference in percentages reflects the size of the corpora. The larger the corpus the more likely some words, particularly grammar words, are repeated, which in turn will reduce the type/token ratio. To demonstrate this, Table 4 and Figure 1 show the accumulative type/token ratios, year after year, for each sister (note that only the letters containing a date are included in this investigation). Taking Julia as an example, the data shows that her first letter in 1884

Table 4: Accumulative type/token ratios for each subcorpus.

JULIA																										
Letter No.	5	6	8	9	10	13	15	17	18	19	20	21	23	26	27	77	30	28	29	32	58					
Date	1884	1884	1888	1889	1889	1890	1890	1890	1891	1891	1891	1891	1892	1893	1893	1893	1893	1894	1894	1895	1927					
Accumulative Type	46	250	335	458	530	592	641	674	724	768	791	823	878	914	942	983	1001	1083	1121	1160	1198					
Accumulative Token	61	600	977	1430	1933	2429	2814	3188	3550	3907	4243	4562	5004	5356	5828	6192	6545	7310	7810	8251	8622					
Ratio%	75.41	41.67	34.29	32.03	27.42	24.37	22.78	21.14	20.39	19.66	18.64	18.04	17.55	17.06	16.16	15.88	15.29	14.82	14.35	14.06	13.89					
ANNIE																										
Letter No.	12	70	22	25	31	33	34	37	38	40	42	43	44	36	46	48	49	52	53	55	56	57	59	60		
Date	1890	1891	1891	1893	1895	1898	1899	1901	1901	1902	1906	1906	1906	1910	1912	1913	1914	1918	1918	1919	1919	1925	1928	1928		
Accumulative Type	205	417	477	611	653	653	721	753	792	817	848	870	891	975	1039	1065	1087	1149	1251	1326	1396	1380	1389	1424		
Accumulative Token	491	1532	2027	2965	3580	4621	5168	5571	6019	6393	6833	7174	7479	8005	8651	9158	9565	10224	11104	11932	12444	12688	12901	13536		
Ratio%	41.75	27.22	23.53	20.61	18.24	14.13	13.95	13.52	13.16	12.78	12.41	12.13	11.91	12.18	12.01	11.63	11.36	11.24	11.27	11.11	11.22	10.88	10.77	10.52		
ALICE																										
Letter No.	7	11	41	45	50																					
Date	1888	1889	1904	1910	1914																					
Accumulative Type	170	305	392	436	464																					
Accumulative Token	395	928	1380	1756	2042																					
Ratio%	43.04	32.87	28.41	24.83	22.72																					
LIZZIE																										
Letter No.	1	2	3	4																						
Date	1876	1876	1876	1877																						
Accumulative Type	409	502	543	631																						
Accumulative Token	1450	2033	2390	2939																						
Ratio%	28.21	24.69	22.72	21.47																						

contains sixty-one tokens (or words) and forty-six types, giving an overall type/token ratio of 75.41%, suggesting that the letter contains a good amount of linguistic diversity with relatively little repetition (although not surprisingly since this is a very short letter). However, looking at the accumulative figures for letters sent between 1884 and 1889 (1,933 tokens and 530 types), there is a much lower type/token ratio (27.42%), which would suggest that some repetition is occurring in the five letters sent during this period. This is to be expected: as mentioned earlier, the larger the corpus the more likely it is that words will be repeated; however, the formulaic nature of letter writing perhaps also goes some way to explaining the dramatic drop in type/token ratio from Julia's first letter sent in 1884, which has a type/token ratio of 75.41%, to her last letter sent in 1927, which has a type/token ratio of 13.89%. The extent of the formulaic writing would require further investigation. Certainly, the openings and closing are likely to follow a standard format, but it would be interesting to examine the body of the letters to see whether they too adopt a set pattern, with less new information being presented over time.

Michael Toolan suggests that what is potentially very interesting when examining accumulative type/token ratios are any sharp 'spikes' or 'dips' in the predictable decline in type/token ratios.[31] In the Lough data there is a sharp decline, or dip, between Julia's first letter sent in 1884 (75.4%) and her second letter sent later that year (41.7%) – a difference of 33.7%. Similarly for Annie, there is a noticeable difference of 14.6% between her first letter sent in 1890 (41.8%) and her second letter sent in 1891 (27.2%), after which the decline is much less pronounced. These dips could be explained, in part, by the length of the letters: in both cases the first letter is quite short whereas the second letter is much longer; however, it could also be indicative of a more formulaic writing style adopted by the two younger sisters, possibly indicative of differences in education between the Lough sisters, although further analysis would be needed to test this hypothesis.

An average type/token ratio can also be calculated. This goes some way to resolving the problem of type/token ratios being lower for larger corpora and higher for smaller corpora, and allows data sets of different sizes to be compared. This is done by calculating the type/token ratio for the first 1,000 words of a corpus, then the next 1,000 words, then the next, and so on. Finally, an average is calculated based on these figures. Table 5 shows the average type/token ratio for the LOUGH Corpus and the two reference corpora (FEMALE Ref. and MALE Ref.). The data shows that female authors have a slightly lower type/token ratio (39.97%) compared with male authors (44.86%). The average type/token ratio for the LOUGH Corpus is slightly lower than both reference corpora at 34.02%, which might support earlier observations that the Lough letters (particularly in the case of Julia Lough and Annie Lough) are perhaps more formulaic than one might expect – an observation that is certainly worth further investigation: what is being repeated and what function does this repetition serve?

Words and frequencies

Having established the lexical density of the letters, the next stage is to look at which words are being repeated (or not, as the case may be). Using Antconc it is possible to create wordlists for the whole corpus and each subcorpus. Table 6 shows the twenty most frequently occurring words in the LOUGH Corpus.

Table 5: Average type/token ratios across three corpora.

	Average Type/Token
FEMALE Ref Corpus	39.97
MALE Ref Corpus	44.86
LOUGH Corpus	34.02

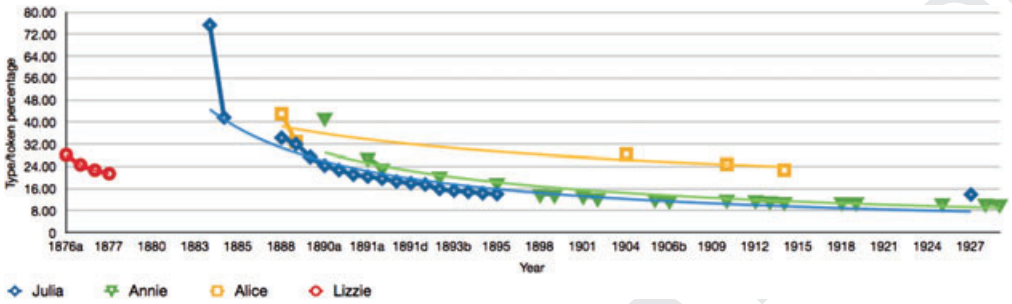

Figure 1: Accumulative type/token ratios for each subcorpus presented visually.

The left column ('Word') shows the words listed in order of frequency with the right column (Raw freq.) providing the actual number of occurrences. Table 6 shows that grammar words are most common: *I* <freq. 1,807>, *you* <freq. 1,324>, *to* <freq. 1,313>, *and* <freq. 1,296>, *the* <freq. 909>. Grammar words are the glue that holds the content together – so it is perhaps not surprising that these words occur more frequently. However, the propensity for certain grammar words over others can be equally revealing. Table 6, for example, shows that the pronouns *I* and *you* are the most frequently occurring words in the LOUGH Corpus with *I* scoring slightly higher than *you*: <freq. 1,807> versus <freq. 1,324> (a ratio of 4:3). One might expect the first person singular pronoun *I* to score high in ego-documents such as letters; however, previous studies have identified gendered variations in terms of pronoun usage. Nicola McLelland, for example, found that female authors tended to refer to themselves using the first person singular pronoun *I* more than male authors; and a study by Arja Nurmi and Minna Palander-Collin found that pronoun usage reflected the power relations between author and recipient – when the relationship was equal (letters between siblings, for example) the first person pronoun usage was high; when the relationship was unequal (letters between children and parents, for example) the first person pronoun usage was low.[32] Their study also found that the sex of the recipient had an effect on pronoun usage, with authors referring to themselves more frequently using *I* when the recipient was female. The current study supports some of these findings with *I* occurring more frequently in the LOUGH Corpus (an average of 47.21 occurrences per 1,000 words) and the FEMALE Corpus (an average of 41.90) than in the MALE Corpus (an average of just 32.94) – see Table 7. The findings did not, however, support Nurmi's and Palander-Collins's observation that first person pronoun usage tends to be greater in letters between authors and recipients of equal status (such as siblings); instead, in the Lough letters, the data showed that *I* occurs slightly more

Table 6: Wordlist for the LOUGH Corpus.

Word	Raw freq.
I	1807
you	1324
to	1313
and	1296
the	909
a	652
is	673
all	586
of	544
it	432
she	418
for	414
her	413
will	411
very	400
in	391
was	385
are	325
have	306
hope	304

Table 7: Occurrence of *I* in each corpus.

	I (Raw freq.)	*I* (Normalised)	*You* (Raw freq.)	*You* (Normalised)
LOUGH Corpus	1807	47.21	1324	34.59
FEMALE Ref.	693	41.90	353	21.34
MALE Ref.	681	32.94	291	14.07

Table 8: Occurrences of *I* in letters sent to parents/siblings.[33]

No. of letters to:	*I* (Raw freq.)	*I* (Normalised)	*You* (Raw freq.)	*You* (Normalised)
Sister (48)	919	45.29	772	38.04
Parents (43)	957	50.37	592	31.16

frequently in letters addressed to the mother (an average of 50.37 occurrences per 1,000 words) than in those addressed to the sister (an average of 45.29 occurrences) – see Table 8. Note that the 'normalised' figures in Table 7 and Table 8 allow meaningful comparisons to be made across data sets of different sizes. It is calculated by dividing the raw frequency by the number of tokens x 1,000. This gives an average frequency (of a particular word or phrase) per 1,000 words.

The pronoun *you* is also potentially very interesting as it has the ability to occupy two grammatical positions (Subject and Object), so its usage might reveal something

Table 9: Position of *I* in projection clauses.

No. of words to the left	L7	L6	L5	L4	L3	L2	L1	**NODE**
Freq. of 'I'	0	3	4	6	27	147	0	**YOU**

about the author/recipient relationship: how the authors are positioning themselves and how they are positioning the recipient. Analysis of the concordance lines for *you* shows that it frequently occurs in the position of Subject of what can be described as a projected clause, where the projecting clause contains the pronoun *I*, as in ***I*** *hope* ***you*** *will write soon* (*I hope* being the projecting clause: the part which projects an idea, fact or proposition; and *you will write soon* being the projected clause: the idea, fact or proposition that is being projected). In short, *you*, in these occurrences, is the real or psychological Subject of these sentences.

Table 9 shows that the pronoun *I* (in these projection clauses) most commonly occurs either two words to the left (L2) <freq. 147> or three words to the left (L3) <freq. 27> of the search word *you*, with the most frequent structures being *I hope you* <freq. 95>, *I suppose you* <freq. 28>, *I am sure you* <freq. 15>, *I wish you* <freq. 11>, *I know/no you* <freq. 11>, *I think you* <freq. 7> and *I am glad you* <freq. 4>). In these instances, the projecting clause (i.e. the clause which introduces the projected clause – the main fact, idea or proposition) contains a mental verb or an adjective carrying epistemic modality (such as *suppose* or *sure* (expressing probability or certainty)), or a mental verb carrying boulemic modality (such as *hope* or *wish* (expressing desire or volition)). It is, arguably, at this point that a phraseological pattern begins to emerge: *I* + *Verb* + *You*; *I* + *BE* + *Adj* + *You*. In any case, the prominence of *you* as doer, agent or focalised, constructed centre of attention, is very striking. I will talk more about projection clauses later in the article.

Breaking the wordlist down further, Table 10 provides the word frequency lists for each subcorpus (as well as the corpus as a whole). The data shows that the grammar words *I* and *you* score high in all four subcorpora; however, although there are <759> occurrences of *I* in the ANNIE Corpus and only <236> occurrences in the LIZZIE Corpus, statistically Lizzie is using *I* much more frequently than the other sisters – on average 67.66 times per 1,000 words, compared with 40.09 for Annie, 50.53 for Julia and 53.53 for Alice. Annie appears to be using *I* (40.09) and *you* (38.66) almost on a 1:1 ratio, perhaps suggesting that she is often directly involving or addressing the recipient in her letters; whereas Lizzie is using *I* (67.66) approximately two and a half times more frequently than she is using *you* (22.94), perhaps suggesting that her letters are more author focused. In all subcorpora the same grammar words (*I*, *you*, *and*, *to*, *the*) are being repeated, which may indicate that certain grammatical structures are also being repeated; this, in turn, may go some way to explaining the low type/token ratio discussed earlier, although further exploration would be needed before any conclusions could be drawn.

Another possible avenue for investigation is the use of *will*, which ranks high across three of the subcorpora: <freq. 229> in the ANNIE Corpus, <freq. 123> in the JULIA Corpus, and <freq. 40> in the ALICE Corpus. The modal verb *will* is interesting as it has several different functions and can be used to express epistemic modality (i.e. certainty/probability), or boulemic modality (i.e. desire/volition). There

Table 10: Wordlist for each subcorpus.

LOUGH Corpus			ANNIE Corpus			JULIA Corpus			LIZZIE Corpus			ALICE Corpus		
Word	Raw freq.	Normalised	Word	Raw freq.	Normalised	Word	Raw freq.	Normalised	Word	Raw freq.	Normalised	Word	Raw freq.	Normalised
I	1807	47.21	and	767	40.51	I	620	50.53	I	236	67.66	I	192	53.53
you	1324	34.59	I	759	40.09	to	422	34.4	to	126	36.12	to	147	40.98
to	1313	34.30	you	732	38.66	and	391	31.87	you	80	22.94	you	124	34.57
and	1296	33.86	to	618	32.64	you	388	31.62	is	73	20.93	and	107	29.83
the	909	23.75	the	508	26.83	the	249	20.3	the	71	20.36	the	81	22.58
a	652	17.03	all	343	18.12	a	208	16.95	she	70	20.07	very	61	17.01
is	673	17.58	a	339	17.91	is	201	16.38	it	50	14.33	all	56	15.61
all	586	15.31	is	311	16.43	all	152	12.39	of	47	13.47	of	54	15.05
of	544	14.21	of	292	15.42	of	151	12.31	an	42	12.04	is	52	14.50
it	432	11.29	it	236	12.47	her	140	11.41	for	40	11.47	a	51	14.22
she	418	10.92	will	229	12.1	am	137	11.17	not	40	11.47	me	43	11.99
for	414	10.82	for	226	11.94	she	134	10.92	her	39	11.18	her	40	11.15
her	413	10.79	very	218	11.51	in	127	10.35	my	38	10.89	will	40	11.15
will	411	10.74	in	210	11.09	will	123	10.03	he	37	10.61	she	39	10.87
very	400	10.45	hope	201	10.62	for	122	9.94	all	35	10.03	they	39	10.87
in	391	10.22	was	201	10.62	not	116	9.45	was	35	10.03	was	38	10.59
was	385	10.06	her	194	10.25	have	113	9.21	but	32	9.17	have	37	10.32
are	325	8.49	are	184	9.72	it	113	9.21	have	32	9.17	in	34	9.48
have	306	7.99	she	175	9.24	was	111	9.05	no	32	9.17	it	33	9.20
hope	304	7.96	well	157	8.29	are	109	8.88	am	31	8.89	but	32	8.92

Table 11: Occurrences of WILL in the LOUGH Corpus.

WILL	Freq.
I will	88
You will	106
she will	30
Maggie, Mary, Lizzie etc. will	24
it will	23
God, heaven will	8
we will	12
lines will	12
they will	11
he will	5
letter will	3

ope they are well so Dear mother I will bring my letter to a close I hope you
I know they are but very few now I will close now dear Sister with best and kindest
very well with it so Der Mother I will conclude with fondest and best love to youe
love to them and now dear Sister I will finish as I cannot wish you a merry Xma
ou know I am thinking of you and I will not forget you next year with Gods help. I
yo will write as soon as you do. I will not write again till I get an answer to
when the last one was not a girl I will not say any more now till I hear from you
isitor as you I dont ask for any I will say good by now with love to you and an
ry you will write soon again and I will send you a longer letter next time and thi
see their grandfather some times I will try and send you their pictures some ti
I was going to write to Mary but I will wait now till I get her next letter y all
t I will have it before Xmas and I will write to you again before Xmas With the

Figure 2: Sample Concordance Lines for 'I WILL'.

are <411> occurrences of *will* in the LOUGH Corpus. Six of these occurrences show *will* functioning as a noun (as in *God's will* and *holy will*), so these can be discounted, leaving <405> occurrences of *will* functioning as a modal verb. Of these <405> occurrences almost half come after the pronouns *I* <freq. 88> and *you* <freq. 106> (see Table 11).

Looking more closely at the concordance lines for *I will* the data shows that in most instances (<61> out of <88> occurrences) *will* is being used in 'signing off' structures to signal the close of the letter (see concordance lines below for examples), with the meaning being one of intention. All of the instances below – *I **will** conclude*, *I **will** finish*, *I **will** bring my letter to a close* – could be substituted with *I intend to* and as such are expressing boulemic modality. As with '*I am writing to you because . . .* ' and '*You ask me X, so I will tell you . . .* ', these meta-discursive phrases help to structure the text as well as serving an interactive function.[34]

Concordance lines display the search term (in this case *I will*) in context. The concordance lines are presented this way (that is with the search term centrally aligned) so as to allow the analyst to notice linguistic patterns – words that typically appear to the right or left of the search term.

The concordance lines for *you will* show that what follows is a limited range of verbs. There are verbs to do with the act of sending/receiving letters (*write*, *send*, *receive*, *get*); and there are verbs to do with cognition (*like*, *forgive*, *excuse*). (See concordance lines below. Note that the verb *keep* is difficult to categorise as it functions

you will		**forgive** me for not writing before now my son
you will		**send** the paper you promised me I
you will		**excuse** all mistakes
you will	→	soon **write** Dearest Mother love to you May john
you will		**keep** to your promise and write again to me
you will		**get** them Dear Mother I will finish up for this
you will		**receive** in due time. the censors are kept very
you will		**like** to reed them let me know if the letters

Figure 3: Sample Concordance Lines for 'YOU WLL'.

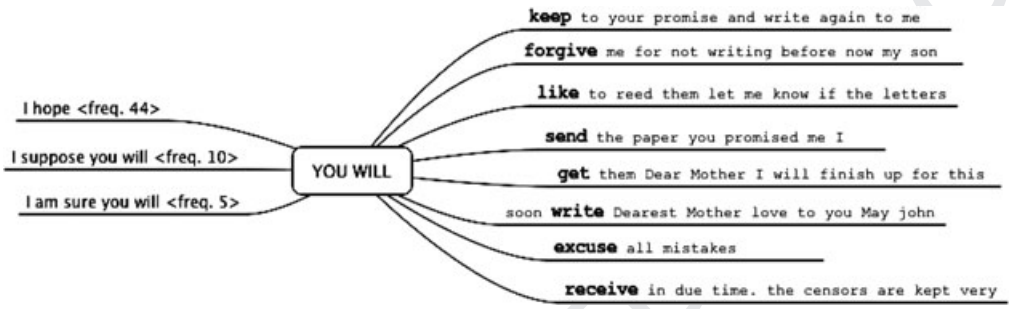

Figure 4: Patterns for 'you will'.

in very different ways and has different meanings depending on the context in which it is being used. In the concordance lines being examined here, *keep* is used in the context of *you will keep to your promise*, where *keep* is part of a fixed expression, meaning 'fulfil your agreement'). In all of these instances of *you will* it is difficult to know whether *will* is expressing epistemic or boulemic modality as there is not enough context for either function to predominate. When, for example, the author says, *you will forgive me for not writing before now*, it is not clear whether *will* is being used to express certainty (as in 'I am quite sure that you will forgive me'), or desire/volition (as in 'I want you to forgive me' or 'I hope you intend to forgive me').

An investigation of the wider context, however, reveals that in most cases (<90> out of <106> occurrences) *you* is the Subject and *will* is the auxiliary modal of a projected clause, preceded by a projecting clause (see Figure 4), with the most frequent patterns being *I hope you will*, *I suppose you will* and *I am sure you will*.

With this wider phraseological context it now becomes more possible to determine the function of *will* in these instances. The type of modality (whether epistemic or boulemic) is projected onto the recipient via the projecting clause, pushing a mild obligation, or placing social pressure onto the addressee to respond in a certain way. The concordance lines for *you will* seem to suggest that *will* is more frequently used to express boulemic modality, with the main pattern (*I hope you will* + *V*) being used to express the author's desire for the recipient's willingness to do something. Through these clauses the author's wants, needs, desires or intentions are transferred onto the recipient – they become the recipient's own and create a psychological bond between both participants.

Another observation which can be made from Table 10 is that across all four subcorpora the only lexical word which appears (in the top twenty) is the verb *hope* with a frequency of <201> in the ANNIE Corpus. However, moving further down the wordlists more content words begin to appear. Table 12 provides a list of the ten most frequent lexical verbs in each subcorpus (i.e. the first ten lexical verbs as they appear, in whatever form, in the wordlist). Note that I am looking at lexical verbs and not auxiliary verbs (HAVE, BE, DO), which tend to serve a grammatical function.

Looking at Table 12 there are two things that stand out. First, there are a high number of mental verbs of cognition, perception and desire (*hope*, *suppose*, *see*, *think*, *like*, *love*, *hear*, *know*), with certain verbs (*hope*, *love* and *think*) appearing across all four subcorpora. Second, nearly all of these verbs appear to be in their base form, with the following exceptions: Lizzie uses the past tense *got* and *thought*, the participle *seen* and *taken* and the continuous form *going*. The high frequency of base forms may in part be explained by the high frequency of *to* and *will* across all four subcorpora (as what tends to follow both *to* and *will* is the base form of the verb, as in *to hear*, *to think*, *will send*, *will go*). However, a closer look at the context surrounding these mental verbs in their base form (of which there are <1,115> occurrences) reveals that they are rarely used after *will* (just <25> instances); they are more frequently used after *to* (a total of <173> instances – the most common structures being *to see* <freq. 72> and *to hear* <freq. 75>); but they are most frequently used in the present tense after the first person singular pronoun *I* (a significant <453> instances).

The high frequency of these mental verbs of cognition, perception and desire is interesting for two main reasons. The first is that these verbs, as explained by Michael Halliday and Christian Matthiessen, 'relate to inner experience (what we experience as going on inside ourselves, in the world of consciousness)' and usually describe emotions, thoughts, or perceptions, thereby providing insight into the psychological worldview of the author.[35] The second is that these verbs are special because they have the ability to project: that is they have the 'ability to set up another clause "outside" the "mental" clause as the representation of the "content" of consciousness'.[36] This latter point appears to support previous findings which show a high frequency of the projecting clause *I hope* (*hope* being a mental verb of desire).

Halliday and Matthiessen make a distinction between the projection of propositions and the projection of proposals, with each type of projection having its own lexicogrammar. 'Whereas propositions, which are exchanges of information [i.e. exchanges which require a verbal response], are projected mentally by processes [verbs] of cognition – thinking, knowing, understanding, wondering, etc. – proposals, which are exchanges of goods-&-services [i.e. exchanges which require a non-verbal response], are projected mentally by processes [verbs] of desire'.[37] Further, what is interesting about the lexicogrammar of proposals is that they can be followed by a future declarative (will + base form) or non-finite (including to-infinitive) dependent clause (as in *I hope you* ***will write*** *soon* or *I hope* ***to hear*** *from you soon*). So, when a verbal response is required the verb is likely to be one of cognition (as in *I know you are trying to do the best you can*). When a non-verbal response is required the verb is likely to be one of desire (as in *I hope you will write often*) – see Table 13. Projection clauses in the LOUGH Corpus will be examined in more detail later in the article.

Table 12: The most frequent lexical verbs in each subcorpus.

LOUGH Corpus			**ANNIE Corpus**			**JULIA Corpus**			**LIZZIE Corpus**			**ALICE Corpus**		
Verb	Freq.	Norm.	Verb	Freq.	Norm.	Verb	Freq.	Norm.	Verb	Freq.	Norm.	Verb	Freq.	Norm.
hope	304	7.94	hope	201	10.62	hope	74	6.03	think	31	8.89	see	23	6.41
write	197	5.15	write	120	6.34	get	66	5.38	get	18	5.16	hope	19	5.30
get	151	3.94	see	74	3.91	hear	65	5.30	write	22	6.31	send	18	5.02
see	143	3.74	love	69	3.64	write	52	4.24	got	11	3.15	get	15	4.18
think	140	3.66	suppose	64	3.38	think	49	3.99	seen	11	3.15	like	15	4.18
love	127	3.32	know	57	3.01	know	42	3.42	thought	11	3.15	love	15	4.18
hear	112	2.93	get	52	2.75	see	40	3.26	hope	10	2.87	know	13	3.62
know	112	2.93	work	48	2.54	love	35	2.85	going	9	2.58	live	13	3.62
suppose	103	2.69	hear	47	2.48	go	34	2.77	love	8	2.29	think	13	3.62
work	84	2.19	think	47	2.48	let	25	2.04	taken	8	2.29	write	13	3.62

Table 13: Examples of propositions and proposals.

Type of Exchange	
Proposition	Proposal
Exchange of information	Exchange of goods & services
Verbal response	Non-verbal response
Mental verbs of cognition (*know*, *think*, *suppose* etc.)	Mental verbs of desire (*hope*, *wish*, *want* etc.)
I know you are trying to do the best you can	*I wish you would write oftener*
I think you are growing smarter	*I hope you will send me the paper*
I suppose you are always busy	*I want you to give five shillings of mine to Mary*

Words in context: n-grams and clusters

Some of the observations above begin to piece together when the next test is carried out, by looking at n-grams. N-grams are: X number of words which appear consecutively Y number of times. The analyst can set the parameters, so, for example, using the n-gram function within Antconc the analyst could search for all 3-grams (three words appearing consecutively) which occur five times or more in the corpus. Table 14 gives a summary of the most frequently occurring 2, 3, 4, 5 and 6-grams.

Table 14 shows that the quantitative findings discussed in previous sections are partly realised in these n-grams, with the lexical verb *hope* followed by the modal auxiliary verb *will* ranking high across four of the five searches. However, this only reveals part of the picture – there are <274> instances of the 2-gram *I hope*, but only <58> instances of *I hope you will*. To get a fuller understanding of the phraseology surrounding a particular word or phrase Antconc has the capability to search for clusters. Figure 5 shows the three, four, five and six word clusters surrounding the phrase *I hope*.

The word tree presented in Figure 5 highlights the lexical and grammatical (or lexicogrammatical) patterns surrounding *I hope*.[38] The diagram shows that not all options are equally probable. What most frequently follows *I hope*, in the Lough corpus, is the pronoun *you*; and what most frequently follows *I hope you* is the modal auxiliary *will*. As the tree branches out the lexicogrammatical choices become fewer, so *I hope you will not*, for example, occurs just twice in the corpus and in itself is not overly significant. Looking at the broader picture, however, through examining lots of evidence at the same time, a phraseological pattern for *I hope* begins to emerge. The question then would be whether this pattern is typical of this data set only, or typical of letters/personal narratives more generally? Is this phraseology used more by one author than another? Finally, what do these linguistic choices reveal about the author, their sex or their experiences?

The quantitative observations so far have teased out several possible lines of inquiry, which could be examined qualitatively. The analyst might, for example, investigate the low type/token ratio and whether or not the significant dip in ratio between Julia's and Annie's first and second letters (indicative of words and phrases being repeated) is typical or unusual amongst different authors (perhaps looking at female/male authors, or authors from different socioeconomic backgrounds). Are, for

Table 14: Most frequent n-grams in the LOUGH Corpus.

2-gram	Freq.	3-gram	Freq.	4-gram	Freq.	5-gram	Freq.	6-gram	Freq.
I am	284	**I hope you**	95	**I hope you will**	58	few lines will find you	12	these few lines will find you	10
I hope	274	**hope you will**	61	love to you and	22	**I hope you will write**	11	**I hope you will write soon**	9
and I	149	I am sure	56	I am sending you	21	with love to you and	11	**hope you will write soon and**	8
to you	147	and I hope	45	and I hope you	19	**and I hope you will**	10	and let me know all the	7
you will	121	Dear Mother I	45	and let me know	17	these few lines will find	10	let me know all the newes	7
all the	120	love to you	42	give my love to	16	and let me know all	9	love to you and John and	7
hope you	114	I am glad	41	Dear Mother I am	15	**hope you will write soon**	9	few lines will find you all	6
I was	108	let me know	40	I am sure you	15	I am sending you some	8	**hope these few lines will find**	6
to her	108	are all well	35	to hear from you	15	let me know all the	8	**I hope these few lines will**	6

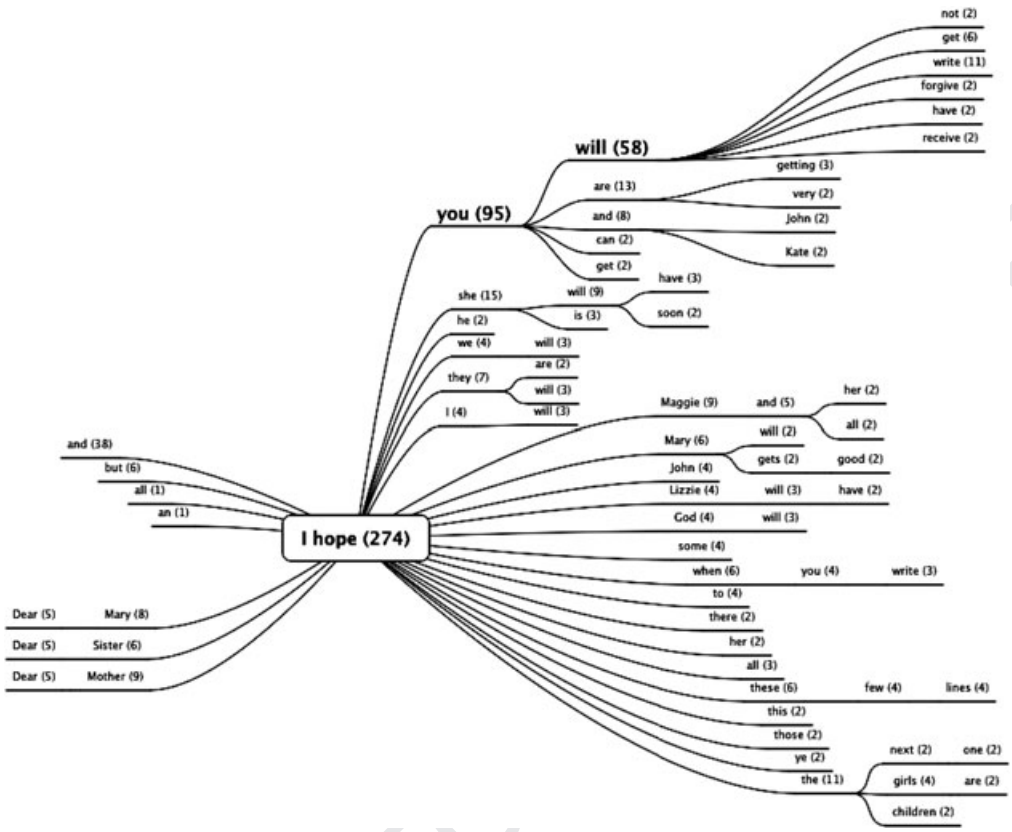

Figure 5: Three-, four-, five- and six-word clusters for 'I hope'.

example, some authors more formulaic than others? To what extent is the main body of the letter formulaic? Which lexicogrammatical structures are being repeated and can any trends be identified? Another line of inquiry might be to examine the high frequency of *I*/*you* in the Lough letters and whether this is in some way genre indicative. Would a study of other text types (narratives, diaries or spoken language) reveal similar findings? As observed by Nicola McLelland, and supported in this study, *I* is more characteristic of female authors; however, how is it being employed, in which context, and when talking about what?[39] Its use in projection clauses (as discussed in this article) is only part of the picture. Alternatively, the analyst might choose to investigate the high frequency of *will* and whether it is functioning in an epistemic or boulemic sense (to show probability, or desire/volition). What other linguistic strategies are used to express modality? Are there any gender or class differences in the use of modality? Arja Nurmi and Minna Palander-Collins, for example, found little variation in modal usage according to social differences; however, they did find some differences in usage between male and female authors, with the modals *will* and *would* being more typical of female writers.[40] Closer investigation showed, however, that these findings varied depending on the author/recipient relationship.

For the present study, I am going to look in more detail at the high frequency of mental verbs of cognition and desire, which occur after the pronoun *I* as part of a

```
ting you know all the News I think I keep  you Well posted. if I did not Write but
my love to them all when you write I mail  you some papers every week hope you get
for her is Kate with her in Galway I sent  you some Transcripts two weeks ago when
d] we are very well at present and I thank you very much for them nice post cards
a letter from you in answer to one I wrote you the first week in September. I hope
```

Figure 6: Examples of Non-Projecting Structures.

projecting clause. I am especially interested in exploring the function of these clauses and what they might reveal about the author/recipient relationships in the Lough letters.

From quantitative to qualitative: concordance lines

I have chosen to explore projection clauses further as the quantitative findings so far appear to suggest that these structures (or phraseological patterns) are frequently used by the Lough sisters, which may be indicative of a local grammar. The main pattern under investigation is: *I* + *V* + *you* + *(modal/aux)* + *V* (as in *I hope you will write*). I have several questions to explore: which verbs (other than *hope*) most commonly occur in projecting clauses; are there more projections of propositions (requiring a verbal response – typically expressed through a cognitive verb), or are there more projections of proposals (requiring a non-verbal response – typically expressed through a verb of desire); which auxiliary verbs most commonly follow *you* in the projected clause; does this pattern (*I* + *V* + *you* + *(modal/aux)* + *V*) attract similar text types – is it genre-indicative; is this pattern used equally by all three sisters, or does one sister use it more than the others, and finally, is this phraseology used as frequently by male and/or other female authors.

I began by carrying out a search on *I* * *you* ('*' is a wildcard meaning 'any word which appears in X position') in the LOUGH Corpus. As the findings in Table 5 illustrate, the search brought up <188> instances of this structure. There are three things to note at this stage. First, this search did not bring up all projection clauses, but only those where *I* occurs one word to the left of the wildcard '*'. As shown in Table 9, previously, *I* can sometimes occur several words to the left of the pronoun *you*, as in ***I*** *hope when you write again* ***you*** *. . . ;* however, for this investigation I focused only on those (most common) structures where *I* occurs directly to the left of the mental verb. Second, the search produced only those projection clauses containing the pronouns *I*/*you* (separate searches would need to be carried out to identify clauses containing *I* + *you*/*he*/*she*/*they*, etc.). Third, not all instances of *I* * *you* are projection clauses. In <41> out of the <188> occurrences *you* is the Object of the main clause (rather than the Subject of a projected clause).

After having removed the non-projecting structures, there are <147> occurrences of *I* ∗ *you* functioning as projection clauses in the LOUGH Corpus.

The auxiliary modals that most frequently follow *you* are listed in Table 16. The data shows that *will* is by far the most common modal used in this structure.

The verbs in Table 15 can be categorised in terms of the experience they are construing. For example, *assure*, *tell*, *thank* and *told* could be described as communicating or saying verbs; *dream*, *hope*, *know*, *like*, *see*, *suppose*, *think*, *want*, *wish* and *wonder* could be described as mental verbs of cognition, perception or desire; and *keep*, *mail*,

Table 15: Search results for *I * you* in the LOUGH Corpus.

I * you	Freq.
assure	4
dreamed	1
hope	95
keep	1
knew	1
know / no	11
like	1
mail	1
received	4
see	2
send	1
sent	5
suppose	28
tell	1
thank	3
think	7
told	1
want	5
wish	11
wonder	1
write	1
wrote	3
TOTAL	**188**

Table 16: Auxiliary modals following *I * you* in the LOUGH Corpus.

	Modal V.	Raw freq.
I * You	can	3
	could	4
	must	1
	ought	1
	will	42

receive, *send*, *sent*, *write*, *wrote* could be categorised as verbs of action. The data shows that the pattern *I* + *Verb* + *You* seems to attract more mental verbs, with *hope*, *know*/*no*, *suppose* and *wish* being the most common.

Of the <147> occurrences of *I* ∗ *you* functioning as a projection clause, the most common verb to occur in this pattern is *hope*. As shown in Figure 5 earlier, over half of all instances of *I hope you* (<58> out of <95>) are followed by *will*. In these instances the author is placing a mild obligation on the recipient to do something – usually write, or forgive for lack of communication. Of the remaining occurrences of *I hope you*, most are standard, formulaic phrases which one might expect in any letter (*I hope you are well*, *I hope you get good health*, *I hope you can read my writing*).

pe you are getting along good and I Know you are trying to do the best you can I
pe you are very well yourself and I Know you are trying to do the best you can I
every as I would wish for you and I know you are doing the best you can in your
must try and keep well if you can I know you never can stop thinking of Dear Annie
haps sooner than you think [sic] I know you would grow young again [Page Four]
not get this before Christmas and I know you would not be happy if you did not
I did not write till the last for I knew you would worry and I was sure you would

Figure 7: Sample Concordance Lines for 'I KNOW YOU'.

y glad I made the change although I think you and mother did not like it by your
r she is going to write this week I think you are growing smarter all the time to
er in your own dear hand writing. I think you done just splendid It was a very nice
r if those things of mine fit her I think you have been more than generous to give
I get her next letter [--damaged] I think you two ought to be very comfortable
y cold weather Julie wrote to you I think you will have hers first I suppose she
to see such style when I go home I think it is nonsense I think you ought to burn

Figure 8: Sample Concordance Lines for 'I THINK YOU'.

ve some very hansom under clothes I wish you could see them evry stitch of clothes
ible cold and talk of snow drifts I wish you could see some of them this last week
und the skirt 20 yds all together I wish you could see it I think I will send you a
ie will come to see you often and I wish you could go see her sometimes give my
icture in my next letter if I can I Wish you would try an have some of your picturs
letter and glad you are all well I wish you would write oftener but I suppose you
ve but I take pleasure in sewing. I wish you was near so I could help you you

Figure 9: Sample Concordance Lines for 'I WISH YOU'.

s 16 years she has only two girls I suppose you all read about our presidents death
would wish so much to see you all I suppose you all felt bad for Parnell it was too
mas how is the winter over there I suppose you are bussey getting ready for xmis
ends here are very well and Alice I suppose you are going to school and is at home
Mary gets good health Dear Mother I suppose you were worried some about that letter
that ye well spend a happy xmas I suppose you will be getting good xmas presents

Figure 10: Sample Concordance Lines for 'I SUPPOSE YOU'.

These, very formulaic, projection structures are commonly found in the openings and closings of letters (as also noted by Dossena) and are described by Mike Scott and Chris Tribble as channel maintainers, helping to sustain the lines of communication between author and recipient.[41]

The Figures below show sample concordance lines for the other main projection clauses.

Looking at the concordance lines for *I know you* and *I think you*, first of all, it appears that *know* and *think* in these clauses are being used as subjective modality markers, rather than true mental projection verbs. These phrases seem to be used when expressing sympathy, or as a way of showing solidarity. The author, in these lines, places themselves in the position of the recipient, imagining their behaviour, what they are doing and how they are feeling. In the case of *I wish* (specifically, *I wish you could see* <freq. 3>) this empathy is reversed and the recipient is invited to imagine something from the author's perspective. Other instances of *I wish* are used to admonish – *I wish you would write oftener* and *I wish you would try to have your photo taken*.

Whereas *wish* is being used to express boulemic modality (the author, in these instances, is expressing a desire for the recipient to do something (*write oftener*) or

Table 17:

	Freq. of I * You	Normalised
Annie	104	5.49
Julia	57	4.65
Lizzie	10	2.87
Alice	17	4.74
LOUGH	188	**4.91**
FEMALE Ref.	35	**2.12**
MALE Ref.	23	**1.11**

experience something (*see her*)), *suppose*, on the other hand, is being used to express epistemic modality. With a degree of certainty, albeit hedged, the author is predicting what the recipient is thinking, feeling or doing. The use of epistemic modality, in these occurrences, emphasises, strengthens and reinforces familial bonds – bonds that are based on past, shared experiences between the two participants. In saying *I suppose you were worried some about that letter* the author is doing more than empathising – they are showing a connection with the recipient which is based on previous and existing knowledge between the two correspondents, which transcends space and time – the message being: 'based on past experiences, and knowing you in the intimate way that I do, my guess is that you are feeling worried'.

The type of projection taking place in these concordance lines (except for instances of *wish*) is a proposition, where the mental verb is one of cognition (*know*, *think*, *suppose*). These projections of propositions require a verbal response, placing a mild obligation on the recipient to (verbally) acknowledge and address the points being raised. These clauses, then, help to facilitate the interactive nature of the letter – establishing and maintaining a dialogue between the two participants. However, as discussed earlier in this article, the most frequently occurring verb in the pattern *I * you* is *hope*, often used to project a proposal (i.e. something which requires a non-verbal response, as in *I hope you will try and be very happy and enjoy yourself*). A closer look at the distributional trends of these *I * you* structures shows that whereas *I hope you* more typically appears in the openings and closing of the letters, *I think*/*know*/*suppose you* tends to occur more frequently in the main body.

Having carried out a search on the projection clause *I * you* in the LOUGH Corpus I then carried out the same search, but this time looking at each subcorpus to see whether one sister uses this structure more than others. The same search was also carried out using the MALE and FEMALE reference corpora to see whether any gender differences (concerning the use of projection clauses) could be identified. The findings are shown in Table 17. Looking at the normalised figures, the data suggests that there is no significant difference in the usage of this structure between Annie, Julia and Alice, although Lizzie seems to use *I * you* much less than her siblings. The data also suggests that female authors use this structure more than male authors; however this is a very general and tentative finding as both reference corpora contain a mixture of authors from different socioeconomic backgrounds, making it difficult to draw any specific conclusions. Indeed, the same search, but this time using a much larger

Table 18:

Corpus	I+1,2you+will	I+1,2you+can	I+1,2you+could
usephem	35.4	4	2.6
brephem	16.6	2.4	1.1
usspok	**13.3**	**58.3**	**18.8**
brbooks	7.4	7.5	4.8
brspok	**4.6**	**53.5**	**27.7**
usbooks	4.5	6.4	3.7
sunnow	3.8	4.1	2.9
strathy	3.1	3.8	2.6
brmags	2.4	4.7	2.2
indy	1.9	2.7	1.2
npr	1.7	11.3	6
usacad	1.6	1.6	0
guard	1.5	2.7	1.4
times	1.5	2.8	1.4
oznews	1.3	2.2	2.2
newsci	1	1.1	0.5
bbc	17	1.6	0.6
usnews	9	1	0.8
wbe	5	0.8	0.5
econ	5	0.3	0.1

(450 million word), contemporary reference corpus (the Bank of English),[42] showed that this structure most commonly occurs in spoken language (see Table 18 – 'brspok' refers to the British spoken language subsection of the corpus and 'usspok' refers to the US spoken language subsection), which could mean that the differences in usage of *I * you* are more indicative of differences in educational background, with letters that adopt a more colloquial, speech-like style making greater use of projection clauses.

Discussion and conclusions

At the beginning of this article I proposed a method of inquiry based on the theory and techniques of corpus linguistics. Taking simple frequency data as the starting point, I was alerted to certain linguistic patterns, which an ordinary reading of the letters may not have allowed. The language contained within the letters was first taken out of its context; it was reorganised to reveal recurring linguistic features worth further, more qualitative, investigation. The findings were then considered within the situational and cultural context of international migration to try and build a picture of how, through letters, family bonds were changed and maintained over space and time.

The approach this article adopts starts with individual words and then examines how those words behave in sentences. What emerges is a specific phraseological pattern ($I + V + you + (modal/aux) + V$), which, further comparative investigations seem to suggest, is used more by female authors than male authors. These projecting structures place the recipient (*you*) – in this case, usually the mother or sister, Mary – as the Subject of the projected clause. However, at the same time, they also place the author

(or more specifically the author's expectations, needs or desires) in the sentence initial position. In other words, these structures lead with some expectation of the author that is highlighted before we reach the main point of the sentence, which requires action, whether verbal or non-verbal, on the part of the recipient. The function of these clauses is to project an imagined narrative onto those back home, arguably serving to maintain a psychological link between the emigrant and their family in Ireland. It is through these, somewhat mundane, repeated phraseological patterns that familial relationships are strengthened and reinforced.

The approach taken in this article is very selective. As mentioned earlier, the initial quantitative investigations highlighted several possible lines of inquiry; however, I chose to follow just one of those, whilst ignoring others. What this approach does offer, however, is a clear, data-led rationale for choosing to examine certain linguistic features in the first place. The numbers themselves are not problematic, nor, necessarily, are the statistical measures or tests that are applied. What, arguably, is problematic are the research questions that are asked in the first place, the data that is used to explore those questions, and/or the conclusions that are later inferred from the results. McLelland's study, discussed earlier, shows how statistics cannot be taken at face value, but should be tested, re-tested and tested again in different ways, against different data sets and by scholars from different disciplinary perspectives.[43] Each line of inquiry will provide different findings, but combined will allow for a fuller, more complete profiling of the female experience of migration. The present study found that a certain pattern appears to be used more by female authors; however until this finding is tested against other data sets (taking into account factors such as social class, educational background, frequency of writing and so on) it is difficult to speak conclusively about the results. Nevertheless, the methodology proposed here is transparent and replicable. The results can be tested, challenged, rejected or confirmed and it is through this process that nuances relating to gender history can begin to emerge.

In many ways this article has brought up more questions than it has provided answers, but one of its main aims was to demonstrate how quantitative methods of analysis might tease out interesting linguistics features for further (quantitative *and* qualitative) analysis. This article has put forward a complementary methodology for examining gender history. It has highlighted some of the possibilities and challenges of using quantitative methods to support, build-on or challenge more qualitative research. Equally, however, it is hoped that some of the quantitative findings discussed here will be taken up by scholars using more qualitative approaches, providing new layers of meaning to the quantitative findings, and, ultimately, the individual emigrants who these numbers represent.

Notes

1. See Kerby A. Miller with David N. Doyle and Patricia Kelleher, 'For Love and Liberty: Irish Women, Migration and Domesticity in Ireland and America, 1815–1920', in P. O'Sullivan (ed.), *The Irish World Wide* (Leicester: Leicester University Press, 1995), pp. 54–61.
2. David A. Gerber, *Authors of Their Lives: The Personal Correspondence of British Immigrants to North America in the Nineteenth Century* (New York: New York University Press, 2006), p. 31.
3. William I. Thomas and Florian Znaniecki, *The Polish Peasant in America,* 5 vols (New York: Dover Publications Inc., 1918–20); Charlotte Erickson, *Invisible Immigrants: The Adaptation of English and Scottish Immigrants in Nineteenth-Century America* (London: The London School of Economics and Political Sciences, 1972); Kerby A. Miller, *Emigrants and Exiles: Ireland and the Irish Exodus to North*

America (Oxford: Oxford University Press, 1985) and Walter D. Kamphoefner, Wolfgang Helbich and Ulricke Sommer (eds), *News from the Land of Freedom: German Immigrants Write Home* (Ithaca and London: Cornell University Press, 1988).

4. Gerber, *Authors of Their Lives*, p. 32.
5. Sonia Cancian, *Families, Lovers and their Letters: Italian Postwar Migration to Canada* (Manitoba: University of Manitoba Press, 2010); Kathleen A. DeHaan, 'Negotiating the Transnational Moment: Immigrant Letters as Performance of a Diasporic Identity', *National Identities* 12 (2010), pp. 107–31; Anna De Fina, and Kendall A. King, 'Language Problem or Language Conflict? Narratives of Immigrant Women's Experiences in the US', *Discourse Studies* 13 (2011), pp. 163–88.
6. Stephan Elspaß, 'Standard German in the 19th-century? (Counter-) Evidence from the Private Correspondence of "Ordinary People"', in Andrew R. Linn and Nicola McLelland (eds), *Standardization: Studies from the Germanic Languages* (Amsterdam: Benjamins, 2002), pp. 43–65; Nicola McLelland, '"Doch mein Mann möchte doch mal wissen . . . " A Discourse Analysis of 19th-century Emigrant Men and Women's Private Correspondence', in Stephan Elspaß, Nils Langer, Joachim Scharloth and Wim Vandenbussche (eds), *Germanic Language Histories from Below (1700–2000)* (Berlin: De Gruyter, 2007), pp. 45–68; Marina Dossena, '"As This Leaves Me at Present" – Formulaic Usage, Politeness and Social Proximity in Nineteenth-Century Scottish Emigrants' Letters', in Elspaß, Langer, Scharloth and Vandenbussche (eds), *Germanic Language Histories from Below*, pp. 1–30; Marina Dossena, '"Many Strange and Peculiar Affairs": Description, Narration and Evaluation in Scottish Emigrants' Letters of the Nineteenth Century', *Scottish Language* 27 (2008), pp. 1–18; Arja Nurmi and Minna Palander-Collin, 'Letters as a Text Type: Interacting in Writing', in Marina Dossena and Ingrid Tieken-Boon van Ostade (eds), *Studies in Late Modern English Correspondence: Methodology and Data* (Bern: Peter Lang, 2008), pp. 21–50.
7. The term 'phraseology', in this paper, refers to the way in which a word typically behaves in context – both the grammatical position it tends to adopt and the words it tends to collocate with. Phraseology and collocation are linked, but whereas collocation tends to refer to word pairings, phraseology refers to extended patterns, where meaning might be carried over several words. The regularities (or patterning) and subtleties in the usage of a word, Hunston argues, are 'difficult to intuit, and [are] observable only when a lot of evidence is seen together', Susan Hunston, *Corpora in Applied Linguistics* (Cambridge: Cambridge University Press, 2002), p. 12. Idioms are fixed expressions and linked to metaphor and figures of speech, but could also be described as phraseology – whereas idioms tend to be fixed and self-contained, phraseology is more open to subtle variations.
8. Dossena '"As This Leaves Me at Present"'.
9. For full details of the design and contents of the corpus see: Marina Dossena, 'Towards a Corpus of Nineteenth-century Scottish Correspondence', *Linguistica e Filologia*. 18 (2004), pp. 195–214.
10. Dossena, '"As This Leaves Me at Present", p. 21.
11. McLelland, '"Doch mein Mann möchte doch mal wissen . . . "'.
12. Studies include: Victoria L. Bergvall, Janet M. Bing, and Alice F. Freed (eds), *Rethinking Language and Gender Research* (London: Longman, 1996); Jennifer Cheshire and Peter Trudgill (eds), *The Sociolinguistics Reader, vol. 2: Gender and Discourse* (London: Arnold, 1998); Janet Holmes, *Women, Men and Politeness* (London: Longman, 1995); Helga Kotthoff and Ruth Wodak (eds), *Communicating Gender in Context* (Amsterdam: Benjamins, 1997); Mary M. Talbot, *Language and Gender*: *An Introduction* (Oxford: Blackwell, 1998); Ruth Wodak, *Gender and Discourse* (London: Sage Publications, 1997).
13. McLelland, '"Doch mein Mann möchte doch mal wissen . . . "', p. 46.
14. McLelland, '"Doch mein Mann möchte doch mal wissen . . . "', p. 55.
15. See studies by Elspaß, and Nurmi and Palander-Collin, previously mentioned. Elspaß, 'Standard German in the 19th-century?'; Nurmi and Palander-Collin 'Letters as a Text Type: Interacting in Writing'.
16. Hunston, *Corpora in Applied Linguistics*, p. 3. The epistemological assumptions that underpin corpus linguistics as a methodology were also discussed by Professor Guy Cook at the 2011 Sinclair Open Lecture at the University of Birmingham, UK.
17. Paul Baker, *Using Corpora in Discourse Analysis* (London: Continuum, 2006).
18. Hunston, *Corpora in Applied Linguistics*, p. 2.
19. Hunston, *Corpora in Applied Linguistics*, p. 28.
20. Hunston, *Corpora in Applied Linguistics*, p. 26.
21. Hunston, *Corpora in Applied Linguistics*, p. 2.
22. See especially Miller, Doyle and Kelleher, 'For Love and Liberty: Irish Women, Migration and Domesticity in Ireland and American, 1815–1920' and Miller, *Emigrants and Exiles: Ireland and the Irish Exodus to North America*.

23. Arnold Schrier, *Ireland and the Irish Emigration, 1850–1900* (Minneapolis: University of Minnesota Press, 1958).
24. Michael Stubbs, 'Conrad in the computer: examples of quantitative stylistic methods', in Ronald Carter and Peter Stockwell (eds), *The Language and Literature Reader* (London: Routledge, 2008), pp. 230–43.
25. Laurence Anthony, 'AntConc Version 3.2.2' (Tokyo, Japan: Waseda University, 2011). <http://www.antlab.sci.waseda.ac.jp/>.
26. Paul Rayson, 'Wmatrix' (Lancaster University, 2009). <http://ucrel.lancs.ac.uk/wmatrix/>.
27. Adam Kilgarriff, Pavel Rychly, P. Smrz. and D. Tugwell, 'The Sketch Engine. Proc. EURALEX 2004', (Lorient, France), pp. 105–16. <http://www.sketchengine.co.uk>.
28. Mike Scott, *WordSmith Tools Version 4* (Oxford: Oxford University Press, 2004).
29. Chris Greaves, 'Concgram.' (John Benjamins Publishing Company, 2005). <http://benjamins.com/#catalog/softwares/cls.1>.
30. Note: Antconc does not distinguish between word class (unless the data is tagged for Parts of Speech), so HOME, whether it was used as a noun or an adjective, would be categorised as one 'type'.
31. Michael J. Toolan, *Narrative Progression in the Short Story* (Amsterdam: John Benjamins Publishing Company, 2009).
32. McLelland, '"Doch mein Mann möchte doch mal wissen . . . "'; Nurmi and Palander-Collin 'Letters as a Text Type: Interacting in Writing'.
33. Note: all letters were included in this investigation (including letters where the authorship is unknown), provided the letter was specifically addressed to either 'mother', 'mother/father' or 'sister'.
34. For more information on metalanguage and metadiscursive phrases see: James Paul Gee, *Social Linguistics and Literacies: Ideology in Discourse* (London: Taylor & Francis, 2008); Annelie Ädel, *Metadiscourse in L1 and L2 English* (Amsterdam: John Benjamins Publishing Company, 2006).
35. Micheal A. K. Halliday and Christian Matthiessen, *An Introduction to Functional Grammar* (London: Arnold, 2004), p. 170.
36. Halliday and Matthiessen, *An Introduction to Functional Grammar*, p. 206.
37. Halliday and Matthiessen, *An Introduction to Functional Grammar*, p. 461.
38. The term lexicogrammar suggests that lexis and grammar cannot be separated, but are instead two ends of the same cline.
39. McLelland, '"Doch mein Mann möchte doch mal wissen . . . "'.
40. Nurmi and Palander-Collin 'Letters as a Text Type: Interacting in Writing'.
41. Dossena, '"As This Leaves Me at Present"'; Mike Scott and Chris Tribble, *Textual Patterns: Key Words and Corpus Analysis in Language Education* (Amsterdam: John Benjamins Publishing Company, 2006).
42. Bank of English (COBUILD and The University of Birmingham, 1991 [2002]). <http://www.titania.bham.ac.uk/docs/svenguide.html>.
43. McLelland, '"Doch mein Mann möchte doch mal wissen . . . "'.

MARKED PROOF

Please correct and return this set

Please use the proof correction marks shown below for all alterations and corrections. If you wish to return your proof by fax you should ensure that all amendments are written clearly in dark ink and are made well within the page margins.

Instruction to printer	*Textual mark*	*Marginal mark*
Leave unchanged	··· under matter to remain	
Insert in text the matter indicated in the margin		New matter followed by or
Delete	through single character, rule or underline or through all characters to be deleted	or
Substitute character or substitute part of one or more word(s)	through letter or through characters	new character or new characters
Change to italics	under matter to be changed	
Change to capitals	under matter to be changed	
Change to small capitals	under matter to be changed	
Change to bold type	under matter to be changed	
Change to bold italic	under matter to be changed	
Change to lower case	Encircle matter to be changed	
Change italic to upright type	(As above)	
Change bold to non-bold type	(As above)	
Insert 'superior' character	through character or where required	or under character e.g. or
Insert 'inferior' character	(As above)	over character e.g.
Insert full stop	(As above)	
Insert comma	(As above)	
Insert single quotation marks	(As above)	or and/or or
Insert double quotation marks	(As above)	or and/or or
Insert hyphen	(As above)	
Start new paragraph		
No new paragraph		
Transpose		
Close up	linking characters	
Insert or substitute space between characters or words	through character or where required	
Reduce space between characters or words	between characters or words affected	